Domain- and Task-Specific Transfer Learning For Medical Segmentation Tasks

Zoetmulder, R.; Gavves, E.; Caan, M.; Marquering, H.

DOI
10.1016/j.cmpb.2021.106539

Publication date
2022

Document Version
Final published version

Published in
Computer Methods and Programs in Biomedicine

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Zoetmulder, R., Gavves, E., Caan, M., & Marquering, H. (2022). Domain- and Task-Specific Transfer Learning For Medical Segmentation Tasks. *Computer Methods and Programs in Biomedicine*, *214*, [106539]. https://doi.org/10.1016/j.cmpb.2021.106539

Computer Methods and Programs in Biomedicine 214 (2022) 106539

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Domain- and task-specific transfer learning for medical segmentation tasks

Riaan Zoetmulder[a,b,*], Efstratios Gavves[b], Matthan Caan[a], Henk Marquering[a,c]

[a] Biomedical Engineering and Physics, Amsterdam UMC, Location AMC, Meibergdreef 15, 1105 AZ Amsterdam, the Netherlands
[b] University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands
[c] Radiology & Nuclear Medicine, Amsterdam UMC, Location AMC, Meibergdreef 15, 1105 AZ Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 25 April 2021
Revised 25 October 2021
Accepted 14 November 2021

Keywords:
Task transfer learning
MRI
Domain adaptation
Deep learning

ABSTRACT

Background and objectives: Transfer learning is a valuable approach to perform medical image segmentation in settings with limited cases available for training convolutional neural networks (CNN). Both the source task and the source domain influence transfer learning performance on a given target medical image segmentation task. This study aims to assess transfer learning-based medical segmentation task performance for various source task and domain combinations. *Methods:* CNNs were pre-trained on classification, segmentation, and self-supervised tasks on two domains: natural images and T1 brain MRI. Next, these CNNs were fine-tuned on three target T1 brain MRI segmentation tasks: stroke lesion, MS lesions, and brain anatomy segmentation. In all experiments, the CNN architecture and transfer learning strategy were the same. The segmentation accuracy on all target tasks was evaluated using the mIOU or Dice coefficients. The detection accuracy was evaluated for the stroke and MS lesion target tasks only. *Results:* CNNs pre-trained on a segmentation task on the same domain as the target tasks resulted in higher or similar segmentation accuracy compared to other source task and domain combinations. Pre-training a CNN on ImageNet resulted in a comparable, but not consistently higher lesion detection rate, despite the amount of training data used being 10 times larger. *Conclusions:* This study suggests that optimal transfer learning for medical segmentation is achieved with a similar task and domain for pre-training. As a result, CNNs can be effectively pre-trained on smaller datasets by selecting a source domain and task similar to the target domain and task.

1. Introduction

Convolutional Neural Networks (CNN) have become the standard approach for medical image segmentation [1]. Accurate CNN-based segmentation approaches typically require a large amount of manually annotated data for training. However, manual annotation of medical images is commonly a time-consuming task, which may require specialized expertise. Reducing the demand for large annotated datasets is therefore an active area of research.

In this study, we focus on transfer learning, which is a broadly applicable strategy to reduce the need for annotated data. Transfer learning aims to reuse a CNN trained on a large dataset rather than directly training a CNN from scratch [2]. In this approach, the weights that are obtained by pre-training are subsequently used to initialize a CNN and perform a different medical image analysis task on a different dataset. The source domain and the source task during pre-training are two relevant aspects of transfer learning [3]. The source domain refers to the type of data used, and the source task refers to the specific application used to pre-train. Analogously, the target domain and target task refer to the type of data and specific application of the main goal. The pre-trained weights can be used on a target task using two strategies: feature extraction or fine-tuning. In feature extraction, the transferred weights are fixed when learning the target task. In fine-tuning, the transferred weights are updated to perform the target task.

Prior studies on transfer learning for medical segmentation target tasks have mostly used two source domains; natural images and medical images. Previous studies that used natural images as the source domain have used two source tasks: ImageNet classification [4] and image segmentation. Two examples with ImageNet classification as the source task are brain-tumor segmenta-

[*] Corresponding author at: Biomedical Engineering and Physics, Amsterdam UMC, Location AMC, Meibergdreef 15, 1105 AZ Amsterdam, the Netherlands.
E-mail address: r.zoetmulder@amsterdamumc.nl (R. Zoetmulder).

https://doi.org/10.1016/j.cmpb.2021.106539
0169-2607/© 2021 Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

tion [5] and stroke lesion segmentation [6]. An example of a study that used natural image segmentation as the source task is colorectal polyp segmentation [7]. The value of pre-training on natural image datasets for medical image analysis target tasks is not clear enough yet. This is because natural image datasets differ from medical image datasets in three important ways. Firstly, medical classification and segmentation tasks often contain a few classes [8,9] whereas natural image classification and segmentation tasks can contain hundreds of different classes [4]. Secondly, natural images are made in heterogeneous settings, whereas medical images are acquired in controlled settings. Hence, the variation in terms of object orientation is larger in natural images than in medical images. Thirdly, natural color images commonly have three channels representing the colors red, green and blue. Whereas scans in medical datasets often do not consist of three channels. This is for example the case if the medical dataset consists of MR or CT scans. Recent work already showed that pre-training on a gray scale version of ImageNet improves transfer learning performance [10]. Transfer learning with medical images as the source domain include segmentation and self-supervised tasks as the source tasks. For example, segmentation source tasks have been used to improve white matter lesions segmentation [11], neonatal brain tissue type segmentation [12], and lung nodule and liver tumor segmentation [13]. Self-supervised source tasks have been used to improve lung nodule segmentation [14].

The choice of the source task has been shown to influence the target task performance. On natural images as the target domain, it has been shown that selecting source tasks that were more similar to their target tasks resulted in better performance on the target task [15]. However, for medical image segmentation target tasks, this has not been established.

If we categorize imaging tasks as self-supervised, classification and segmentation tasks, for the natural image source domain, studies have used classification [5,6] and segmentation [7], but not self-supervised source tasks. Differently, for the medical image source domain, studies have used segmentation [11–13] and self-supervised [14], but not classification source tasks. It can therefore be concluded that the effect of the source domain and tasks on the target medical segmentation accuracy has not yet extensively been evaluated.

In the current study, we empirically investigate the effect of the choice of source task and domain on the performance of multiple medical segmentation target tasks: stroke lesion, MS lesion, and brain anatomy segmentation on T1 MR. Furthermore, we aim to compare the optimal source-target task/domain combination with a common benchmark in transfer learning research: pre-training on ImageNet.

2. Related work

2.1. Transfer learning

The goal of transfer learning is to pre-train a model on a source task and reuse the information the model has learned to improve performance on a target task [16]. Transfer learning was first shown to work in neural networks by Pratt et al. [17] and was subsequently applied to problems in computer vision [18] and medical image analysis [2].

A commonly used approach to apply transfer learning is to pre-train the CNN on a task and domain and to (partially) fine-tune the CNN on a target task and domain. In computer vision, the CNN is often pre-trained on the ILSVRC‘12 (ImageNet) dataset [4]. Work on medical image analysis has used CNNs pre-trained on ImageNet classification [19] as well. Recently, the use of CNNs pre-trained on ImageNet for medical image analysis has been questioned. Recent work has found that the transfer learning benefits gained from pre-training on ImageNet classification were inconsistent on diabetic retinopathy grade classification on fundus photographs and thoracic pathology classification on chest X-Ray scans [20]. As a result, other data sets and tasks have been investigated as alternatives to ImageNet classification for transfer learning in medical image analysis [13,14].

Other research has developed alternative methodologies to the pre-training and fine-tuning procedure that is widely used [21,22]. Spot Tune is a method that adaptively decides to freeze or fine-tune specific layers in the CNN for each input image [21]. Co-Tuning is a method that fully re-uses the pre-trained CNN by learning a mapping from the target classes to the source classes and uses these labels as an additional supervision signal during fine-tuning.

2.2. Domain adaptation

Domain adaptation (or *transductive transfer learning*) is a special case of transfer learning in which the source and target task are the same but the data distribution of the source and target domains differ [3]. The goal of domain adaptation is to build domain invariant models that learn similar features from the source and target domains. Techniques are based on minimizing the difference between the feature distributions acquired from the source and target domain [23–28]. For example, prior work has proposed a method by which statistical dependence was preserved by using a reproducing kernel Hilbert space [26]. Other work has proposed a manifold criterion to create an intermediate domain, which is related to the target domain, using source data [27].

Research in medical image analysis has also made use of domain adaptation [11,29,30]. Prior research on MR has applied domain adaptation to generalize automated segmentation of white matter hyper intensities to follow-up scans using fine-tuning [11]. Another method has used adversarial learning to generalize segmentation of abnormalities on brain MR scans after traumatic brain injury [29].

2.3. Task transfer learning

Task transfer learning (or *inductive transfer learning*) is a special case of transfer learning in which the source and target task differ but the data distribution of the source and target domains is the same [3]. In computer vision, several studies have investigated the relationship between different tasks [15,31,32]. One study investigated the relationship between individual source and target tasks to create a taxonomy of the degree to which tasks transfer to each other by fine-tuning each target task on each source task [15]. This study also investigated the performance gain achieved when features from models pre-trained on different source tasks were combined to learn a target task. Developing a practical method to decide which source tasks are the most important in decision making support when computational resources are limited has consequently been investigated [33]. Other research has focused avoiding having to fine-tune networks by finding the affinity between various classification tasks [32].

2.4. Few shot learning

Few-shot learning is a machine learning sub-field that aims to learn from a few training examples (for example five cases) per individual class [34]. In computer vision, much research has been dedicated to developing few-shot learning methods [35–37]. For medical image analysis, few-shot learning has also been adopted for organ segmentation [38,39].

3. Materials

This section describes the used datasets, the pre-processing, and the CNN architectures.

3.1. Datasets

Two natural image datasets were used to pre-train the CNNs: the **taskonomy** dataset [15], consisting of 4.6 million images of indoor scenes with multiple annotations per image, and the **ImageNet** dataset [4] consisting of 1.2 million images of 1000 different objects.

Four medical datasets that consist of T1 brain MRIs were used in our study. Firstly, we have used the Brain-Age Healthy Cohort **(BAHC)** [40] to pre-train the CNNs. The BAHC is a dataset compiled of 2001 scans from 14 different data sources. Ground truth annotations were created using a combination of Nipype [41], FSL [42] and ITK [43]. Additional information about participants and the scan acquisition parameters can be found in table 3, in appendix 9.1. Secondly, we have used the ATLAS R1.2 **(Stroke lesion dataset)** [44], which is a manually annotated T1 MRI dataset of lesions after ischemic stroke and consists of scans from 304 patients. Each scan contains at least one lesion. The annotations consist of a primary lesion and secondary non-contiguous lesions. Additional information about the scans is included in table 4, appendix 9.1. Thirdly, we have included 30 scans of the **Multiple Sclerosis (MS) lesion** dataset [45]. The scans were acquired using a 3T Siemens Magnetom Trio. The resulting scans had a resolution of 0.57 × 0.57 × 3.00 – 3.30 mm. Fourthly, we included the **Brain Anatomy (BA) dataset** which consists of 35 scans from the OASIS project [46]. Manual annotations were combined into six classes [47]. The images were acquired on a Siemens Vision 1.5T scanner and had a resolution of 1.0 × 1.0 × 1.25 mm.

3.2. Pre-processing

All scans were reoriented and resampled to the MNI-ICBM 152 template [48], axially zero padded to the taskonomy dataset dimensions. Empty axial slices were discarded, being 61, 29, 11 and 11 slices starting from the top of the scan volume for the BAHC, stroke lesion, MS lesion, and BA dataset respectively. For the MS lesion and BA dataset also 50 slices from the bottom were discarded. Voxel intensities above the 99th and below the 1st percentile were clipped, and intensities were then normalized using min-max normalization.

3.3. CNN Architectures

This study uses CNNs developed in earlier work [15]. The CNN architectures consist of an encoder (Fig. 1A) along with only a decoder (Fig. 1B), an encoder along with a decoder and a discriminator, (Fig. 1C) or an encoder along with a fully connected layer (Fig. 1D). The CNNs were pre-trained on the source tasks.

3.3.1. The encoder

The encoder is based on the ResNet-50 [49] architecture. To ensure that the encoder has a latent space of 16 × 16 × 4, the fully connected layers of the ResNet-50 are replaced by a transposed convolution with a stride of two. The encoder uses the ReLU activation function [50].

3.3.2. The decoder

The decoder consists of convolutions and transposed convolutions. It up-samples the features from 16 × 16 × 4 to an image of 256 × 256 with the number of channels required for the task. Each convolution layer had a stride of one and each transposed convolution had a stride of two. The kernel size was 3 × 3. The first two layers were convolutional layers, the eight subsequent layers alternated between a transposed convolutional layer and regular convolutional layer, ending with the latter. The decoder used a leaky ReLU activation function [50] with alpha set to 0.2.

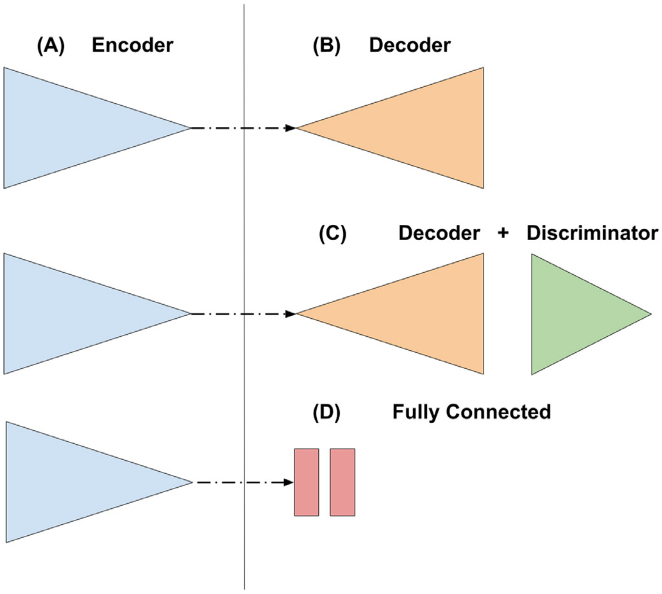

Fig. 1. Components of the CNN architectures. A CNN architecture consists of an encoder (A) followed by either a decoder (B), a decoder and a discriminator (C), or a fully connected layer (D).

3.3.3. The discriminator

The image and the decoder output were used as input of the discriminator [51]. The first convolutions had a stride of one and a kernel size of five. Next, two convolutional layers followed with a stride of four. The final two layers had a kernel size of four and a stride of one. The discriminator used a leaky ReLU activation function [50] with alpha set to 0.2.

3.3.4. The fully connected block

The fully connected block consisted of two fully connected layers. The first layer had a hidden size of 2048, the second layer a size of 16. The first fully connected layer [52] uses a ReLU activation function[50] and the second fully connected layer uses a softmax activation function.

4. Methods: source and target tasks

The source and target domains have been addressed in Section 3. In the current section, the source and target tasks are addressed.

4.1. Source tasks

A schematic representation of the source tasks of the equal (T1 MRI) and unequal (natural image) domain is shown in Fig. 2. The hyper-parameters used to pre-train the CNNs on the equal domain as the target tasks are described in appendix 9.2. The CNNs were pre-trained on axial slices from the scans. The hyperparameters used to pre-train the CNNs on the unequal domain are described in appendix 9.3.

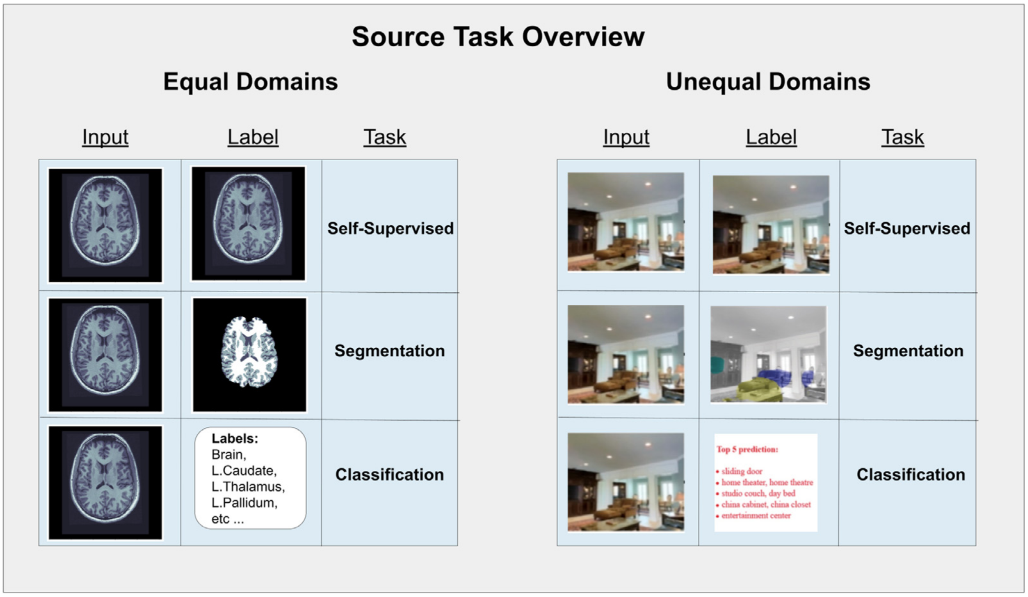

Fig. 2. Schematic representation of the source tasks for the equal (T1 MRI) (left) and unequal (natural images) (right) domain. For each task, an example is given of the input (left column), the label (middle column), and the task name (right column). The source tasks are; self-supervised, segmentation, and classification.

4.1.1. Segmentation source tasks

The segmentation source tasks were brain tissue and indoor object segmentation for the equal and unequal domain, respectively. The brain tissue ground truth for three classes was generated using the FAST algorithm [53] from the FSL toolkit. The indoor object ground truth was available for 17 classes [54] & [15]. The segmentation CNN architecture consisted of the encoder, followed by the decoder. The loss function used was the weighted cross-entropy and used class weights calculated by [55].

4.1.2. Self-Supervised source task

For both source domains, auto-encoding was the self-supervised source task [56]. The auto-encoding CNN architecture consisted of an encoder, a decoder and a discriminator. The loss consisted of the weighed sum of the L_1 norm and the GAN loss. The weights used were 0.996 for the L_1 norm and 0.004 for the GAN loss.

4.1.3. Classification source tasks

The equal domain classification source task was brain and sub-cortical structure classification, which was a multi-label classification task. Each axial slice contained annotations indicating whether the brain and specific sub-cortical structures were present. Ground truth segmentations of these structures were generated using the FIRST [57] and BET [58] algorithm from the FLS toolkit. The classification of a subset of 100 ImageNet classes of indoor scenes was used as the unequal domain object classification task [15]. The binary cross entropy was used as the loss function class-wise. The used CNN architecture was the encoder followed by a fully connected block.

4.2. Comparison to pre-training on the full-extent of ImageNet

The previously described transfer learning experiments include source tasks with a similar amount of data to pre-train for a fair comparison. The most commonly used source task, ImageNet classification [4], uses at least ten times more data than the models described above. To compare the other approaches to the most commonly used benchmark, a CNN pre-trained on the full-extent of the ImageNet classification dataset was included as a source task.

4.3. Target tasks

For all experiments, the encoder was initialized using one of the source tasks, and the decoder was initialized randomly. The CNNS were fine-tuned with multiple sub-sample sizes, which will be referred to as the **fine-tuning set size**. The CNNs were fine-tuned on axial slices obtained from the included scans.

All CNNs used a batch size of 32, a learning rate of 10^{-4} and a weight decay of $2 \cdot 10^{-4}$. The number of epochs varied per task and are discussed per experiment. After half the training epochs were completed, the learning rate was decayed by 10. For the segmentation tasks, the weighted cross entropy was calculated voxel wise.

4.3.1. Stroke lesion segmentation

The stroke lesion segmentation target task consisted of segmenting stroke lesions from non affected tissue and background.

The data was split randomly into a training and testing set of 200 and 104 scans, respectively. The fine-tuning set size was incremented from 10 to 100 scans with steps of 10. For each fine-tuning set size, ten fine-tuning sets were randomly sampled from the training set. CNNs were fine-tuned for 30 epochs.

4.3.2. Multiple sclerosis lesion segmentation

The MS lesion segmentation task consisted of segmenting MS lesions from non affected tissue and background. The MS lesion data was split randomly into a fine-tuning set of 20 scans and a testing set of 10 scans, respectively. CNNs were fine-tuned for 60 epochs.

4.3.3. Brain anatomy segmentation

The brain anatomy segmentation task consisted of segmenting seven anatomical regions. The BA dataset was divided into 15 scans for fine-tuning and 20 for testing [47]. CNNs were fine-tuned for 300 epochs.

4.4. Evaluation metrics

The Dice coefficient and the mean intersection over union (mIOU) were used to assess the spatial accuracy for single and

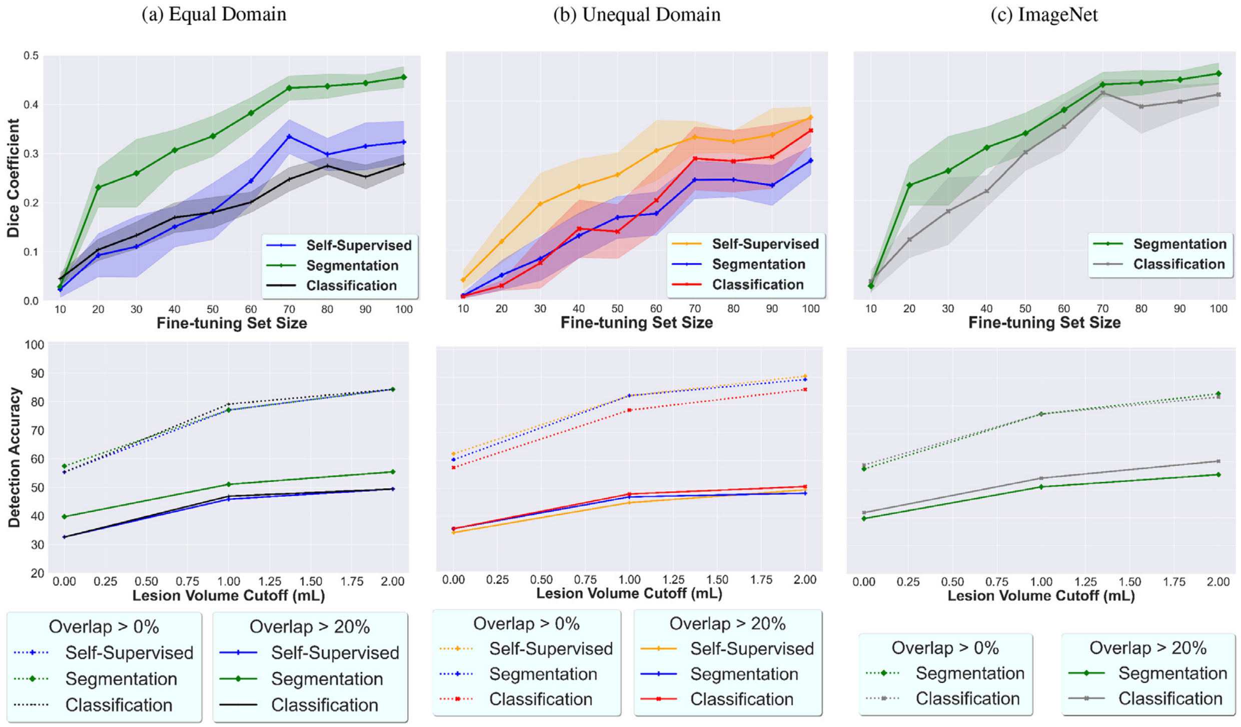

Fig. 3. Spatial agreement (top) and lesion detection accuracy (bottom) for the stroke lesion segmentation task. The spatial agreement is assessed using the Dice coefficient as a function of the fine-tuning set size. The lesion detection accuracy is given as a function of the lesion volume cutoff, which is the minimal volume for which lesions are considered detected. The lesion detection accuracy is calculated with 0% and 20% overlap cutoffs (bottom). (a) Equal domain (T1-MRI) pre-training. Included source tasks are self-supervised, segmentation, and classification. (b) Unequal domain (natural images) pre-training. Included source tasks are self-supervised, classification, and segmentation. (c) Comparison of full-extent ImageNet classification pre-training and pre-training on equal domain and task: brain tissue segmentation.

multi-class segmentation tasks, respectively. Both measures were calculated per axial slice and averaged over the entire test set.

For the MS and stroke lesion segmentation tasks, the lesion detection accuracy was assessed. A connected component analysis was performed on the ground truth segmentation mask to separate all individual lesions in the mask. Voxels were considered to be part of a common lesion if they were 8-connected in the mask. A lesion was considered detected if the percentage of accurately automatically quantified voxels exceeded a pre-set threshold. This threshold is referred to as the *voxel overlap cutoff*. We used voxel overlap cutoff values of 0% and 20%. The higher the voxel overlap cutoff, the more difficult it is for a lesion to be detected by the model. In addition, we wanted to assess the lesion detection accuracy for different lesion volumes. To this end, we progressively excluded lesions below a pre-set volume. We refer to this parameter as the *lesion volume cutoff*, and we used pre-set values of 0 mL (including all lesions), 1 mL and 2 mL. For stroke lesion detection, this analysis was conducted on one of the CNNs fine-tuned with a fine-tuning set size of 100.

5. Experiments & results

5.1. Stroke lesion segmentation

The Dice coefficient and lesion detection accuracy for the stroke lesion segmentation models are shown in Fig. 3. For the equal domain experiments (Fig. 3a), the segmentation source task transfer learning model resulted in the largest Dice coefficient. If the voxel overlap cutoff was set to zero, the lesion detection accuracy was similar for models pre-trained on each source task. When the voxel overlap cutoff was set to 20%, the lesion detection accuracy dropped overall. However, the segmentation source task resulted in models with a higher lesion detection accuracy than the other models.

For unequal source domains, shown in Fig. 3b, the self-supervised source task resulted models with the highest Dice coefficient. No single source task resulted in a model with a higher lesion detection accuracy than models pre-trained on the other source tasks.

Comparing equal and unequal domains, the best equal domain source task (segmentation) resulted in models with a higher Dice coefficient than the other source tasks. Models pre-trained on the classification source tasks consistently yielded a low Dice coefficient. Results for the lesion detection accuracy were more ambiguous. With a voxel overlap cutoff greater than 20%, the model pre-trained on the best equal domain source task resulted in a higher lesions detection accuracy. However, a voxel overlap cutoff greater than 0% resulted in the model pre-trained on the unequal domain source tasks achieving a higher lesion detection accuracy.

In comparison to the full-extent ImageNet pre-trained model, the equal source domain and task model obtained the highest Dice coefficient (Fig. 3c). The lesion detection accuracy is similar for both approaches (Fig. 3c). However, when the voxel overlap cutoff was set to 20%, the ImageNet pre-trained model detected more lesions.

5.2. Brain anatomy segmentation

Spatial agreement results for the BA segmentation target task are shown in Table 1. For the equal domain, the results show that the model pre-trained on the segmentation source task resulted in the highest mIOU. For the unequal domain, all models had lower accuracy than the model pre-trained on the best equal domain source task. Even the full-extent ImageNet pre-trained model was

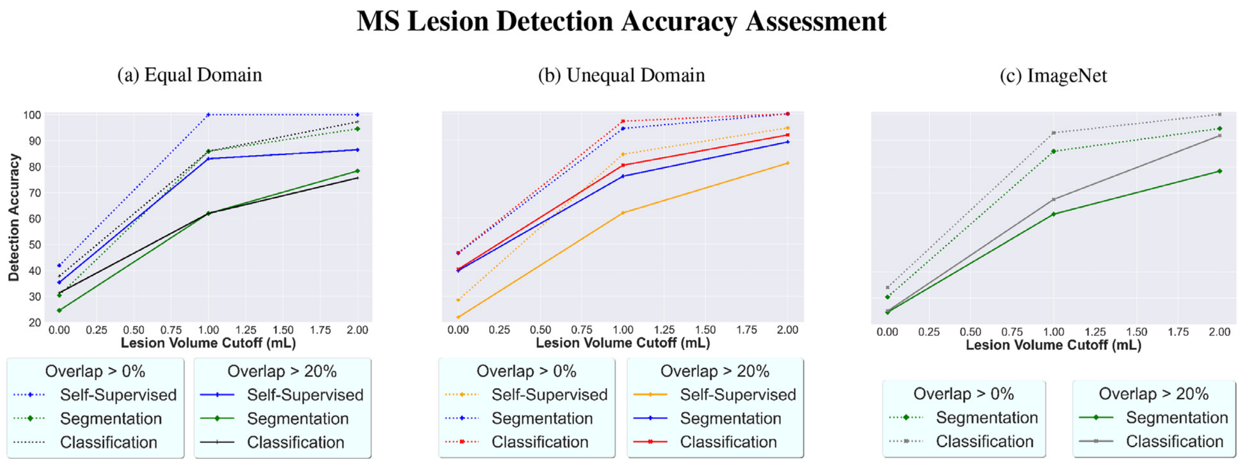

Fig. 4. MS lesion detection accuracy as a function of the lesion volume cutoff with 0% and 20% volume overlap cutoffs. (a) Equal domain pre-training (T1 MRI) for the self-supervised, segmentation, and classification source tasks. (b) Unequal domain (natural images) pre-training for the source tasks classification, and segmentation. (c) Pre-training with the full-extent ImageNet classification source task vs medical image segmentation source task.

Table 1
mIOU of the brain anatomy multi-class segmentation task for various source domains (T1 MRI vs natural images) and source tasks. The self-supervised task is autoencoding. The highest mIOU is underlined.

	Brain Anatomy Segmentation
Equal Domain: T1 MR	
Segmentation	<u>0.62</u>
Self-Supervised	0.58
Classification	0.59
Unequal Domain: Natural Images	
Segmentation	0.56
Self-Supervised	0.57
Classification	0.55
ImageNet Classification	0.52

Table 2
Dice coefficient for the MS lesion segmentation for various source domains. The equal domain is T1 MR, and the unequal domain is natural images. The highest Dice coefficient is underlined. The self-supervised tasks are autoencoding.

	MS Lesion Segmentation
Equal Domain: T1 MR	
Segmentation	0.16
Self-Supervised	0.12
Classification	0.14
Unequal Domain: Natural Images	
Segmentation	0.14
Self-Supervised	0.16
Classification	0.13
ImageNet Classification	<u>0.17</u>

outperformed by the best performing equal domain transfer learning model.

5.3. Multiple sclerosis lesion segmentation

The results for the MS lesion segmentation show that the Dice coefficient was generally low, with segmentation and the self-supervised source task resulting in the highest Dice coefficient for equal and unequal domain, respectively. In addition, the classification source task on the equal and unequal domain consistently results in a low Dice coefficient. The results are shown in Table 2.

The MS lesion detection results are shown in Fig. 4. For the equal domain, the self-supervised source task resulted in models with the highest lesion detection accuracy, regardless of the voxel overlap or lesion volume cutoff.

For the unequal domain, the classification and segmentation source tasks resulted in models with a higher lesion detection accuracy regardless of the lesion volume cutoff. The full-extent ImageNet pre-trained model resulted in a slightly higher Dice coefficient and lesion detection accuracy relative to the best performing equal domain source task.

5.4. Qualitative analysis

We performed a qualitative analysis by visual comparison of the automatically generated segmentation masks. Examples of these masks are shown in Fig. 5.

Visual inspection resulted in two observations. Firstly, pre-training on the segmentation task on an equal domain resulted in the largest segmentations. Secondly, the self-supervised source task pre-trained on the unequal domain resulted in implausible segmentations. For example, it falsely predicted lesions in both hemispheres.

For the BA segmentation, the ImageNet pre-trained model resulted in a larger number of false positives for white matter segmentation. In addition, the ImageNet pre-trained model resulted in a larger number of false positives and negatives of the cerebellum.

6. Discussion

Our study provides the first empirical comparison between the most frequently chosen source tasks for medical segmentation target tasks. These source tasks are self-supervised, classification and segmentation on both natural images and medical images for medical segmentation transfer learning performance. Previous work gave evidence of the advantage of transfer learning for medical segmentation target tasks using various source tasks and domains [11–14]. We build on top of this work by studying how transfer learning performance was influenced by the choice of source task and domain. Our findings corroborate those found in a single source domain [15]; source tasks that are more similar to the target task result in higher transfer learning performance.

A possible explanation of our finding, is that higher layers are more specialized to perform the source task [18]. As such, these layers have a better initialization to perform tasks that are similar to the source task, which may result in finding a better optimum after fine-tuning.

Research comparing the transfer learning performance of various source tasks in medical image analysis has focused mostly

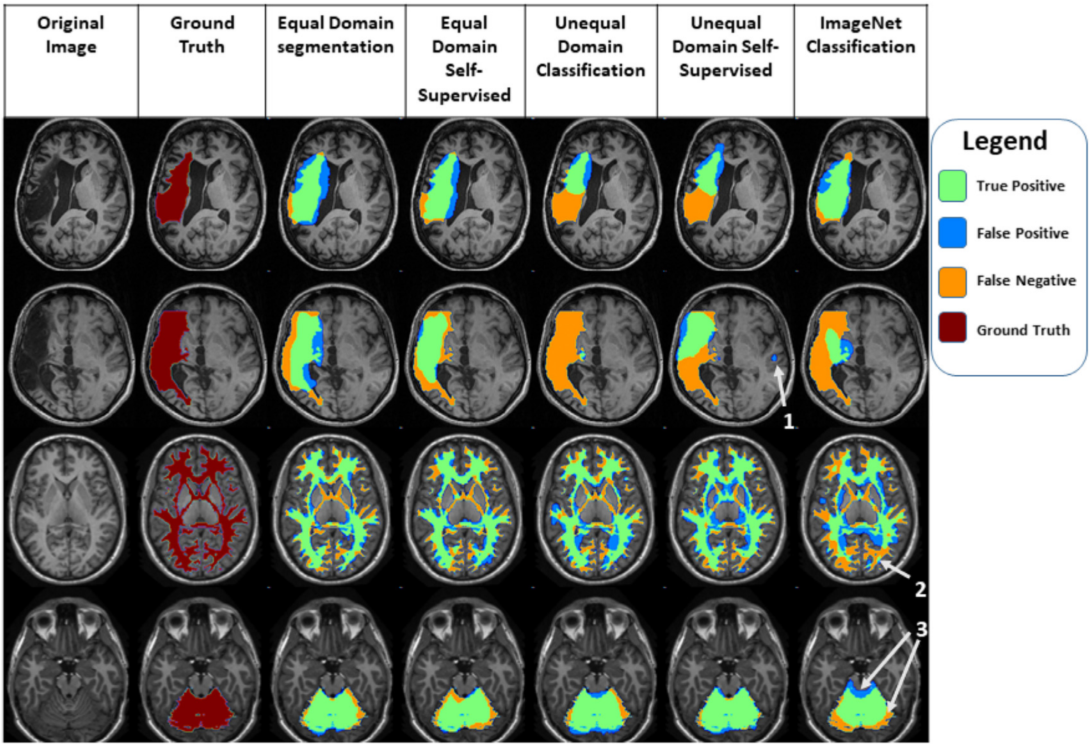

Fig. 5. Examples of segmentation results obtained by the different source tasks on the target tasks stroke lesion (top two rows) and BA (bottom two rows) segmentation. From left to right, the original T1 MR scan, the ground truth segmentation and different segmentation results are shown after transfer learning using the source tasks: segmentation, the self-supervised source task, and classification on the same domain, the self-supervised source task on an unequal domain and full-extent ImageNet pre-training.

on target classification tasks. Previous work showed that supervised and self-supervised pre-training on CT scans yielded a lower transfer learning performance than self-supervised pre-training on natural images [59]. Another study provided evidence that pre-training on a large natural image dataset resulted in an equivalent classification performance using less data, faster convergence, greater robustness against domain shift, and little influence on the calibration of uncertainty estimation [60]. However, other work provided evidence that pre-training on natural images only resulted in faster convergence of the networks but did not result in an improvement of classification performance [20].

Studies done on natural images have similar findings to our own; similar domains and tasks result in optimal transfer learning performance. This was shown for various classification target tasks [32] and a segmentation target task [61]. However, a domain shift can cause the most similar source task to the target segmentation task to result in sub-optimal transfer learning performance [61]. This result is corroborated by another study, which showed that increasing the amount of data used for pre-training could adversely influence transfer learning performance on classification target tasks if the additional data was not from a similar domain [62]. Our study observes the same phenomenon in transfer learning for medical image segmentation.

To promote comparability, our work firstly focused on a single target domain, allowing us to exploit large amounts of data available to pre-train CNNs and the existence of open source tools to automatically create segmentation annotations. Secondly, we only used the ResNet-50 architecture, because pre-trained weights were available for natural image analysis tasks. The specialized U-Net [63] architecture for medical image analysis tasks does not have weights from pre-training tasks on natural images available. Furthermore, U-Net does not generalize well to classification tasks in a straightforward manner because of skip connections. Thirdly, we chose fine-tuning and not feature extraction by pre-trained weights. Hereby, we limited the influence of otherwise confounding variables on the accuracy assessments of transfer learning.

A first limitation of our approach is that we have only tested two source domains, natural images and T1 MR brain scans, and one target domain, T1 MR brain scans. However, there is evidence from other studies that pre-training a model on an equal domain to the target task results in similar better results [13,59]. Hence, it is reasonable to assume this would apply to other medical image domains as well.

A second limitation of our research is the focus on 2D segmentation. Methods that use 3D self-supervised source tasks result in better performance than 2D ImageNet pre-trained models [14]. There is prior evidence showing that transfer learning strategies work well across different architectures [13,64,65]. Therefore, we expect our findings in 2D to generalize to 3D as well.

A third limitation is that the Dice coefficient was low for all MS lesion segmentation target task regardless of the pre-training approach. In T1-weighed MRI, MS lesions in white matter appear as slightly hypo-intense, with intensities similar to gray matter. This makes segmenting MS lesions a challenging task. We found that transfer learning is of limited additional benefit for this challenging task.

A fourth limitation in our study is that we have used a single CNN architecture, i.e. ResNet-50. This architecture was used because pre-trained weights were available for all necessary natural image tasks. Several studies have shown that transfer learning strategies work well across different architectures [13,64,65]. Considering the significant amount of additional computational resources needed to pre-train additional models and fine-tune them, in our study we focus on an archetypal CNN architecture to derive our insights.

In this work, we have thoroughly compared the medical segmentation performance for various target segmentation tasks on brain MR imaging using transfer learning with various source domains and tasks used for pre-training. Our results suggest that medical segmentation tasks benefit from transfer learning with pre-training on segmentation source tasks on the same domain.

7. Conclusion

Our transfer learning experiments targeting brain MRI segmentation tasks suggest that selecting a similar (segmentation) source task and domain results in equal or better spatial agreement than other choices of source task and domain combinations. Even with a source dataset 10 times as large, pre-training on ImageNet classification did not outperform the equal source and target task and domain combination in two out of three target tasks: stroke lesion and brain anatomy segmentation. However, source task and domain selection have an inconsistent effect on the lesion detection accuracy.

Declaration of Competing Interest

This publication is based on the STAIRS project under the TKI-PPP program. The collaboration project is co-funded by the PPP Allowance made available by Health-Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. Dr. H.A. Marquering is a cofounder and shareholder of Nico.Lab. Dr. M.W.A. Caan is shareholder of Nico.Lab.

Acknowledgments

This publication is based on the STAIRS project under the TKI-PPP program. The collaboration project is co-funded by the PPP Allowance made available by Health-Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. Dr. H.A. Marquering is a cofounder and shareholder of Nico.Lab. This work has been supported by the donation of a Quadro P6000 GPU by NVIDIA.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2021.106539.

References

[1] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imaging 32 (4) (2019) 582–596.

[2] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[3] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[5] J. Amin, M. Sharif, M. Yasmin, T. Saba, M.A. Anjum, S.L. Fernandes, A new approach for brain tumor segmentation and classification based on score level fusion using transfer learning, J. Med. Syst. 43 (11) (2019) 1–16.

[6] T. Han, V.X. Nunes, L.F.D.F. Souza, A.G. Marques, I.C.L. Silva, M.A.A.F. Junior, J. Sun, P.P. Reboucas Filho, Internet of medical things-based on deep learning techniques for segmentation of lung and stroke regions in CT scans, IEEE Access 8 (2020) 71117–71135.

[7] J. Kang, J. Gwak, Ensemble of instance segmentation models for polyp segmentation in colonoscopy images, IEEE Access 7 (2019) 26440–26447.

[8] M.A. Akhloufi, M. Chetoui, Chest XR COVID-19 Detection, 2021. Online; accessed September 2021, https://cxr-covid19.grand-challenge.org/)

[9] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (brats), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[10] Y. Xie, D. Richmond, Pre-training on grayscale imagenet improves medical image classification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, p. 0.

[11] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C.R. Guttmann, F.-E. de Leeuw, C.M. Tempany, B. Van Ginneken, et al., Transfer learning for domain adaptation in MRI: application in brain lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 516–524.

[12] G. Zeng, G. Zheng, Multi-stream 3D FCN with multi-scale deep supervision for multi-modality isointense infant brain mr image segmentation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 136–140.

[13] S. Chen, K. Ma, Y. Zheng, Med3D: transfer learning for 3D medical image analysis, arXiv preprint arXiv:1904.00625(2019).

[14] Z. Zhou, V. Sodha, M.M.R. Siddiquee, R. Feng, N. Tajbakhsh, M.B. Gotway, J. Liang, Models genesis: generic autodidactic models for 3D medical image analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 384–393.

[15] A.R. Zamir, A. Sax, W. Shen, L.J. Guibas, J. Malik, S. Savarese, Taskonomy: disentangling task transfer learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3712–3722.

[16] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.

[17] L.Y. Pratt, et al., Discriminability-based transfer between neural networks, Adv. Neural Inf. Process. Syst. (1993) 204.

[18] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? in: Advances in Neural Information Processing Systems, 2014, pp. 3320–3328.

[19] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. (2020) 104115.

[20] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, Transfusion: understanding transfer learning with applications to medical imaging, arXiv preprint arXiv:1902.07208(2019).

[21] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, R. Feris, SpotTune: transfer learning through adaptive fine-tuning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4805–4814.

[22] K. You, Z. Kou, M. Long, J. Wang, Co-tuning for transfer learning, Adv. Neural Inf. Process. Syst. 33 (2020).

[23] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 97–105.

[24] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: International Conference on Machine Learning, PMLR, 2015, pp. 1180–1189.

[25] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.

[26] S. Wang, L. Zhang, W. Zuo, B. Zhang, Class-specific reconstruction transfer learning for visual recognition across domains, IEEE Trans. Image Process. 29 (2019) 2424–2438.

[27] L. Zhang, S. Wang, G.-B. Huang, W. Zuo, J. Yang, D. Zhang, Manifold criterion guided transfer learning via intermediate domain generation, IEEE Trans. Neural Netw. Learn. Syst. 30 (12) (2019) 3759–3773.

[28] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.

[29] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 597–609.

[30] H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey, arXiv preprint arXiv:2102.09508(2021).

[31] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, S. Savarese, Which tasks should be learned together in multi-task learning? in: International Conference on Machine Learning, PMLR, 2020, pp. 9120–9132.

[32] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C.C. Fowlkes, S. Soatto, P. Perona, Task2Vec: task embedding for meta-learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6430–6439.

[33] Q. Chen, Z. Zheng, C. Hu, D. Wang, F. Liu, On-edge multi-task transfer learning: model and practice with data-driven task allocation, IEEE Trans. Parallel Distrib. Syst. 31 (6) (2019) 1357–1371.

[34] Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: a survey on few-shot learning, ACM Comput. Surv. (CSUR) 53 (3) (2020) 1–34.

[35] L. Zhang, J. Liu, B. Zhang, D. Zhang, C. Zhu, Deep cascade model-based face recognition: when deep-layered learning meets small data, IEEE Trans. Image Process. 29 (2019) 1016–1029.

[36] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.

[37] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, ICML Deep Learning Workshop, vol. 2, Lille, 2015.

[38] A.G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, C. Wachinger, 'Squeeze & excite' guided few-shot segmentation of volumetric images, Med. Image Anal. 59 (2020) 101587.

[39] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, D. Rueckert, Self-supervision with superpixels: training few-shot medical image segmentation without annotation, in: European Conference on Computer Vision, Springer, 2020, pp. 762–780.

[40] J.H. Cole, R.P. Poudel, D. Tsagkrasoulis, M.W. Caan, C. Steves, T.D. Spector, G. Montana, Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker, Neuroimage 163 (2017) 115–124.

[41] K. Gorgolewski, C.D. Burns, C. Madison, D. Clark, Y.O. Halchenko, M.L. Waskom, S.S. Ghosh, Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python, Front Neuroinf. 5 (2011), doi:10.3389/fninf.2011.00013.

[42] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, et al., Advances in functional and structural mr image analysis and implementation as FSL, Neuroimage 23 (2004) S208–S219.

[43] T.S. Yoo, M.J. Ackerman, W.E. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxas, R. Whitaker, Engineering and algorithm design for an image processing Api: a technical report on ITK-the insight toolkit, Stud. Health Technol. Inform. (2002) 586–592.

[44] S.-L. Liew, J.M. Anglin, N.W. Banks, M. Sondag, K.L. Ito, H. Kim, J. Chan, J. Ito, C. Jung, S. Lefebvre, et al., The anatomical tracings of lesions after stroke (ATLAS) dataset-release 1.1, bioRxiv (2017) 179614.

[45] Ž. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, Ž. Špiclin, A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus, Neuroinformatics 16 (1) (2018) 51–63, doi:10.1007/s12021-017-9348-7.

[46] D.S. Marcus, T.H. Wang, J. Parker, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults, J. Cogn. Neurosci. 19 (9) (2007) 1498–1507.

[47] P. Moeskops, J.M. Wolterink, B.H. van der Velden, K.G. Gilhuijs, T. Leiner, M.A. Viergever, I. Išgum, Deep learning for multi-task medical image segmentation in multiple modalities, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 478–486.

[48] V.S. Fonov, A.C. Evans, R.C. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, Neuroimage (47) (2009) S102.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385(2015).

[50] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. ICML, vol. 30, 2013, p. 3.

[51] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[52] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167(2015).

[53] Y. Zhang, M. Brady, S. Smith, Segmentation of brain mr images through a hidden Markov random field model and the expectation-maximization algorithm, IEEE Trans. Med. Imaging 20 (1) (2001) 45–57.

[54] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully convolutional instance-aware semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2359–2367.

[55] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[56] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[57] B. Patenaude, S.M. Smith, D.N. Kennedy, M. Jenkinson, A Bayesian model of shape and appearance for subcortical brain segmentation, Neuroimage 56 (3) (2011) 907–922.

[58] S.M. Smith, Fast robust automated brain extraction, Hum. Brain Mapp. 17 (3) (2002) 143–155.

[59] T. Schlegl, J. Ofner, G. Langs, Unsupervised pre-training across image domains improves lung tissue classification, in: International MICCAI Workshop on Medical Computer Vision, Springer, 2014, pp. 82–93.

[60] B. Mustafa, A. Loh, J. Freyberg, P. MacWilliams, A. Karthikesalingam, N. Houlsby, V. Natarajan, Supervised transfer learning at scale for medical imaging, arXiv preprint arXiv:2101.05913(2021).

[61] K. Dwivedi, G. Roig, Representation similarity analysis for efficient task taxonomy & transfer learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12387–12396.

[62] J.-C. Su, S. Maji, B. Hariharan, When does self-supervision improve few-shot learning? in: European Conference on Computer Vision, Springer, 2020, pp. 645–666.

[63] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[64] S. Kornblith, J. Shlens, Q.V. Le, Do better ImageNet models transfer better? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2661–2671.

[65] M. Christopher, A. Belghith, C. Bowd, J.A. Proudfoot, M.H. Goldbaum, R.N. Weinreb, C.A. Girkin, J.M. Liebmann, L.M. Zangwill, Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs, Sci. Rep. 8 (1) (2018) 1–13.